SORBONNE UNIVERSITÉ

Rapport de projet

Sciences des données

Rayane MAKHLOUF

Amayas SADI

Année universitaire: 2021/22

Description de l'ensemble des expérimentations menées sur la base de données agribalyses ainsi que les résultats obtenus:

Après avoir étudié la base de données agribalyses, nous avons choisi de traiter deux problèmes concernant celle-ci

.

- Un premier problème supervisé qui est celui de la classification d'un aliment selon le groupe d'aliment auquel il appartient. Notre but étant de pouvoir classifier un aliment sans pour autant connaître quel est le type de cet aliment.
- Un deuxième problème non supervisé qui est celui du clustering d'un groupe d'aliments qui permet de spécifier certains aliments d'autres aliments selon leurs impacts environnementaux.

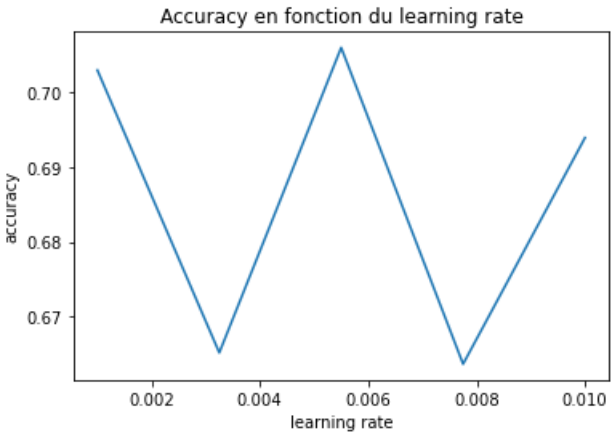
Avant de passer les données à nos classifiers, nous allons d'abord effectuer une normalisation sur les données. Celle-ci est dû à la différence au niveau de l'échelle des valeurs des données. Par la suite, nous allons transformer les données catégorielles en numérique afin de pouvoir les passer à nos classifiers.

Afin de répondre à notre première problématique, nous allons passé nos données normalisées numériques aux classifiers Perceptron et KNN puis au random Forest.

Perceptron:

Afin de déterminer la valeur optimale de learning rate qui maximise la performance de notre classifier Perceptron, nous avons choisis une approche qui consiste en le calcul de la performance de notre classifier perceptron pour des valeurs différentes du learning rate et d'en choisir celle qui maximise cette dernière. Nous avons choisi une méthode de validation croisée.

En analysant le graphique représentant les performances du classifieur Perceptron en fonction des différentes valeurs du learning rate, on constate que la valeur du learning rate optimale est estimée à 0.0055 avec une performance de 70.6%.

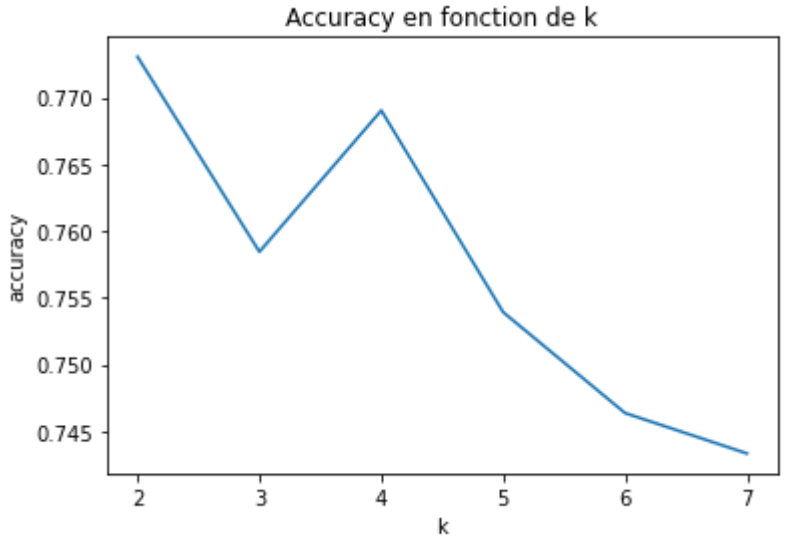


KNN:

Afin de déterminer la valeur optimale de k (nombre de voisins) qui maximise la performance de notre classifieur KNN, nous avons choisis une approche qui consiste en le calcul de la performance de notre classifier pour des valeurs de k différentes et d'en choisir celle qui maximise cette dernière. Nous avons choisi une méthode de validation croisée.

En analysant le graphe représentant les performances du classifieur KNN en fonction des différentes valeurs de k, on constate que la valeur de k optimale est 2 avec une performance de 77.31.

Pour la suite, on choisit le classifieur knn entraîné avec notre valeur optimale de k.

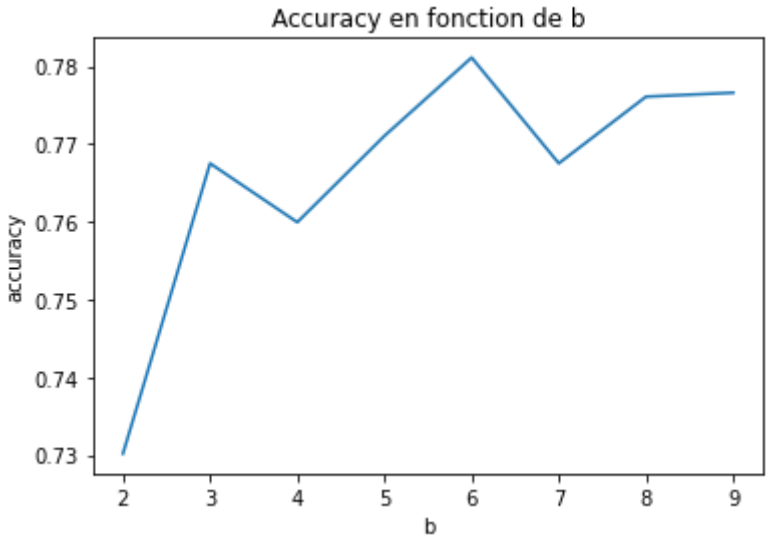


RandomForest:

Afin de déterminer la valeur optimale de b qui maximise la performance de notre classifier RandomForest, nous avons choisis une approche qui consiste en le calcul de la performance de notre classifier pour des valeurs de b différentes et d'en choisir celle qui maximise cette dernière. Nous avons choisi une méthode de validation croisée.

En analysant le graphe représentant les performances du classifieur RandomForest en fonction des différentes valeurs de b, on constate que la valeur de b optimale est 9 avec une performance de 79.12%.

Pour la suite, on choisit le classifier RandomForest entraîné avec notre valeur optimale de b.



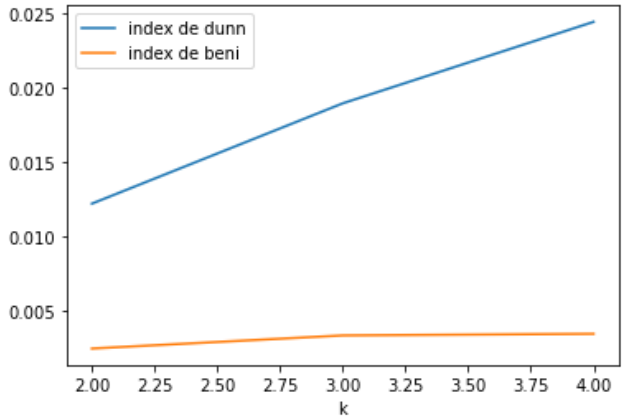
Place maintenant à la deuxième problématique qui est celle du problème du clustering d'un groupe d'aliments (problème non supervisé).

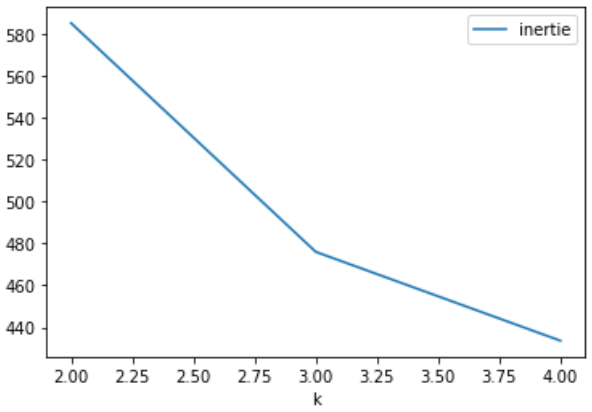
Nous avons choisi le groupe d'aliments de produits céréaliers auquel on va appliquer un clustering après normalisation des données et leurs transformations en données numériques.

Algorithme des K moyennes :

Afin de classer les aliments au sein d'un même groupe selon différents clusters, on utilise l'algorithme des k moyennes qui rend l'ensemble des centroïdes et une matrice d'affectation des aliments.

En utilisant l'index de Dunn et de Beni, on évalue par la suite les performances de notre clustering en fonction des différentes valeurs de k.





En analysant le graphe des indexes de Dunn et de Beni ainsi que celui représentant les inerties globales de l'algorithme des k moyennes en fonction des différentes valeurs de k, on constate que la valeur de k optimale est 2.